A New Person-Fit Statistic for the Lognormal Model for Response Times

Sandip Sinharay, Educational Testing Service

An Updated Version of this document appeared in the Journal of Educational Measurement. The website for the article is https://onlinelibrary.wiley.com/doi/full/10.1111/jedm.12188

The citation for the article is: Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. Journal of Educational Measurement, 55(4), 457-476.

Note: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170026. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education or of Educational Testing Service.

A New Person Fit Statistic for The Lognormal Model for Response Times

Sandip Sinharay, Educational Testing Service

August 10, 2018

Note: Any opinions expressed in this publication are those of the author and not necessarily of Educational Testing Service.

A New Person Fit Statistic for The Lognormal Model for Response Times

Abstract

Response-time models are of increasing interest in educational and psychological testing. This paper focuses on the lognormal model for response times (van der Linden, 2006), which is one of the most popular response-time models, and suggests a simple person-fit statistic for the model. The distribution of the statistic under the null hypothesis of no misfit is proved to be a χ^2 distribution. A simulation study and a real data example demonstrate the usefulness of the suggested statistic.

Key words: χ^2 statistic, time intensity parameter.

Response-time models (RTMs) are psychometric/statistical models that are used to analyze response times. The use of RTMs has been suggested to improve precision of examinee ability estimates (e.g. Bolsinova & Tijmstra, 2018), to detect test fraud (e.g., van der Linden & Guo, 2008), to detect speededness (e.g., Schnipke & Scrams, 1997), to improve test construction (e.g. van der Linden, 2007), and to test substantive theories about cognitive processes (e.g., van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011). Several RTMs have been suggested by, for example, Bolsinova and Tijmstra (2018), Klein Entink, Fox, and van der Linden (2009), Klein Entink, van der Linden, and Fox (2009), Maris (1993), Maris and van der Maas (2012), Rasch (1960), Schnipke and Scrams (1997), Thissen (1983), van der Linden (2006), van der Linden (2007), van der Maas et al. (2011), and Wang and Hanson (2005). Extensive reviews of RTMs include Lee and Chen (2011), Schnipke and Scrams (2002), van der Linden (2009), van der Linden (2016), and van Rijn and Ali (2017).

The specific RTM used in an application may not fit the response times of all the examinees and tge responses of some examinees to some items would be aberrant due to reasons such as item preknowledge, other types of test fraud, and test speededness (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014). Detection of the examinees whose response times are aberrant can be performed using person-fit analysis based on the response times (e.g., Marianti et al., 2014; van der Linden & Guo, 2008), which refers to an evaluation of whether the RTM fits the response times of the individual examinees. This paper suggests a new statistic that can be used to perform person-fit analysis based on the response times. The null distribution of the new statistic is proved to be a χ^2 distribution.

The next section includes a review of a popular RTM, existing approaches for estimation of the parameters of the model, and existing approaches for the assessment of person fit for the model. The Methods section includes the description of the new person-fit statistic and the derivation of its null distribution. The Simulation section includes a comparison of the Type I error rate and the power of the new person-fit statistic to those of an existing statistic. The Real Data section includes an application of the new person-fit statistic to an operational data set. Discussion and conclusions are provided in the last section.

Background

The Lognormal Model for Response Times

Let us consider a test that includes I items. Let t_i denote the response time of a randomly chosen examinee[1] on item i, where $i = 1, 2, \cdots, I$. Let us define

$$y_i = \log(t_i)\cdot$$

According to the lognormal model for response times (LNMRT; van der Linden, 2006), y_i's, $i = 1, 2, \cdots, I$, are independent given τ and

$$y_i|\tau \sim \mathcal{N}\left(\beta_i - \tau, \frac{1}{\alpha_i^2}\right), \tag{1}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. The parameter τ is the examinee's speed parameter; a larger value of the parameter results in smaller expected response times on all the items for the examinee. The parameter β_i is the time-intensity parameter for item i; a larger value of the parameter results in larger expected response times for all examinees on the item. The parameter α_i is the discrimination parameter for item i; a larger value of the parameter leads to more information on and hence smaller standard error of the examinee speed parameters. To ensure model identifiability, the mean of the prior distribution $g(\tau)$ on τ is restricted to be zero (e.g., van der Linden & Guo, 2008).

The LNMRT is arguably one of the most popular RTMs. The model was considered, either to analyze only the response times, or to analyze the response times and item scores, by, for example, Bolsinova and Tijmstra (2018), Boughton, Smith, and Ren (2017), Glas and van der Linden (2010), Qian, Staniewska, Reckase, and Woo (2016), van der Linden (2007), van der Linden (2009), van der Linden (2016), van der Linden and Glas (2010), and van der Linden and Guo (2008). Bolsinova and Tijmstra (2018, p. 13) commented that the LNMRT is used in most applications of RTM.

[1]No subscript is used here for the examinees because the existing statistic and the new statistic will be defined for one randomly chosen examinee.

Note that researchers such as Klein Entink et al. (2009), Marianti et al. (2014), and Fox and Marianti (2017) used a parameterization of the LNMRT that is slightly different from that in Equation 1 and involves the assumption that

$$y_i|\tau \sim \mathcal{N}\left(\phi_i(\beta_i - \tau), \frac{1}{\alpha_i^2}\right).$$

However, the introduction of the parameter ϕ_i was proved to be unnecessary by van der Linden (2016, p. 268).

Estimation of the Model Parameters

A Gibbs sampler (e.g., Gelman, Carlin, Stern, & Rubin, 2003) was suggested by van der Linden (2006) to estimate the item parameters of the LNMRT. That approach has been used in almost all applications of the model and the R package LNIRT (Fox, Entink, & Klotzke, 2017) can be used to implement the Gibbs sampler. Glas and van der Linden (2010) suggested an approach to compute the MLEs of the item parameters when the LNMRT is used along with the three-parameter logistic model (3PLM) to jointly analyze both response times and item scores.

In this paper, the item parameters (α_i^2's and β_i's) are treated as known during the derivation of the null distribution of the person-fit statistic. The assumption of known item parameters is a reasonable assumption if the investigator has calibrated these parameters from a large sample of examinees so that the estimation bias is negligible. In addition, the assumption of known item parameters in deriving distributions of person-fit statistics is common in person-fit analysis based on item scores (e.g., Meijer & Sijtsma, 2001; Snijders, 2001; Sinharay, 2016) and Glas and Dagohoy (2007) found that estimation of item parameters had a negligible effect on the properties of person-fit statistics based on item scores.

van der Linden (2006) showed that given α_i^2's and β_i's, the MLE of the person speed parameter τ can be obtained as

$$\hat{\tau} = \frac{\sum_i \alpha_i^2(\beta_i - y_i)}{\sum_i \alpha_i^2}. \tag{2}$$

Because $\hat{\tau}$ is a linear combination of normal random variables y_i's, it has a normal distribution (because of, for example, Theorem 2.4.1 of Anderson, 1984, p. 25) with mean and variance given by

$$\mathrm{E}(\hat{\tau}) = \frac{\sum_i \alpha_i^2(\beta_i - \beta_i + \tau)}{\sum_i \alpha_i^2} = \frac{\tau \sum_i \alpha_i^2}{\sum_i \alpha_i^2} = \tau \text{ and } \mathrm{Var}(\hat{\tau}) = \frac{1}{\left(\sum_i \alpha_i^2\right)^2} \sum_i \frac{\alpha_i^4}{\alpha_i^2} = \frac{1}{\sum_i \alpha_i^2}. \quad (3)$$

Note that even though the expressions of $\mathrm{E}(\hat{\tau})$ and $\mathrm{Var}(\hat{\tau})$ in Equation 3 are derived conditional on (or, given a fixed value of) τ, the conditioning is not shown in these derivations and in several derivations in the remaining of this paper for convenience.

Existing Person-fit Statistics

Marianti et al. (2014) and Fox and Marianti (2017) suggested the person-fit statistic given by

$$l^t = \sum_i \alpha_i^2(y_i - \beta_i + \tau)^2. \quad (4)$$

Equation 1 implies that

$$\alpha_i(y_i - \beta_i + \tau) = \frac{y_i - E(y_i)}{\sqrt{\mathrm{Var}(y_i)}}$$

is the true standardized residual corresponding to y_i and follows the standard normal distribution; also, given τ, the $\alpha_i(y_i - \beta_i + \tau)$'s are independent over i because of the local independence assumption of the LNMRT. Therefore, given τ, l_t is the sum of squares of I independent standard normal random variables and hence follows the χ^2 distribution with I degrees of freedom (e.g., Anderson, 1984, p. 280). That is, given τ,

$$l^t = \sum_i \alpha_i^2(y_i - \beta_i + \tau)^2 \sim \chi_I^2, \quad (5)$$

where χ_I^2 denotes the χ^2 distribution with I degrees of freedom.

Marianti et al. (2014) recommended the use of a Markov chain Monte Carlo (MCMC) algorithm to fit the LNMRT to the data and recommended the computation of l^t in each iteration of the MCMC algorithm.[2] For each examinee, the proportion of the values of l^t

[2]Each iteration of the MCMC algorithm produces draws of α_i, β_i and τ—these draws can be used to compute l^t in each iteration of the algorithm.

that are larger than the 95th percentile of the χ^2_I distribution is the estimate of the posterior probability of an aberrant response pattern; a person misfit is concluded at 5% level if the estimated posterior probability is larger than 0.95 (Marianti et al., 2014, p. 437).

van der Linden and Guo (2008) suggested a Bayesian approach to determine if the response time of an examinee-item combination is aberrant. They proved that the posterior distribution of the predicted value of the response time on item i conditional on $\mathbf{y}_{-i} = (y_1, y_2, \cdots, y_{i-1}, y_{i+1}, \cdots, y_I)$, is normal. Then, the standardized residual is computed as

$$sr_i = \frac{y_i - E(y_i|\mathbf{y}_{-i})}{\sqrt{\text{Var}(y_i|\mathbf{y}_{-i})}}.$$

If the absolute value of the sr_i is larger than an appropriate quantile of the standard normal distribution, the response time for the examinee for item i is concluded as aberrant. One can compute the sr_i's for all examinee-item combinations and, as in van der Linden and Guo (2008), make a conclusion about the overall fit of the model based on the overall percentage of the sr_i's that are statistically significant. One can also compute the sr_i's for an examinee over all the items and then combine information on the aberrance over all the items for the examinee to assess person fit, as in Boughton et al. (2017, p. 181).[3] The use of this approach to assess person fit depends on the cutoff chosen; van der Linden and Guo (2008) did not recommend a cutoff and it is not clear what cutoff leads to satisfactory Type I error rate and power. Therefore, the approach of van der Linden and Guo (2008) is not considered henceforth to assess person fit. One can use the approach of van der Linden and Guo (2008) to pinpoint the misfit for the examinees (by, for example, finding the items that are leading to aberrant residuals for one examinee or a group of examinees) who were flagged by the method of Marianti et al. (2014) or the new statistic to be described shortly.

As the above review shows, there exists no frequentist approach to person-fit analysis for response-time models. This void is surprising given that there exist frequentist approaches to assess, for example, item fit (e.g., Glas & van der Linden, 2010; Ranger & Ortner, 2012),

[3]In such an approach, misfit is concluded for a person if the number of statistically significant standardized residuals for the person is larger than an appropriate cutoff.

fit of the local independence assumption (Glas & van der Linden, 2010), independence of responses and response times (van der Linden & Glas, 2010), and differential item functioning (Glas & van der Linden, 2010) for response-time models. Therefore, the frequentist approach suggested later in this paper is the first of its kind.

Method: A New Person-fit Statistic

The Statistic

The statistic l_t cannot be used as a person-fit statistic in a frequentist approach because τ is unknown in such an approach. However, a frequentist person-fit statistic can be obtained by replacing τ in l_t by its MLE $\hat{\tau}$.

Let us consider a randomly chosen examinee whose speed parameter is τ. Let's define the estimated standardized residual on item i for the examinee as

$$r_i = \alpha_i(y_i - \beta_i + \hat{\tau})\cdot \tag{6}$$

Because r_i is a linear combination of normal random variables, it follows a normal distribution. Further, because of Equations 1 and 3, $E(r_i)$ can be obtained as

$$E(r_i) = \alpha_i(\beta_i - \tau - \beta_i + \tau) = 0, \tag{7}$$

and, because

$$\text{Cov}(y_i, \hat{\tau}) = \text{Cov}\left(y_i, \frac{\sum_i \alpha_i^2(\beta_i - y_i)}{\sum_i \alpha_i^2}\right) = -\frac{\alpha_i^2}{\sum_i \alpha_i^2}\,\text{Var}(y_i) = -\frac{\alpha_i^2}{\sum_i \alpha_i^2}\frac{1}{\alpha_i^2} = -\frac{1}{\sum_i \alpha_i^2}, \tag{8}$$

The variance $\text{Var}(r_i)$ can be obtained using Equations 1 and 3 as

$$\text{Var}(r_i) = \alpha_i^2\,\text{Var}(y_i + \hat{\tau}) = \alpha_i^2\left(\frac{1}{\alpha_i^2} + \frac{1}{\sum_i \alpha_i^2} - \frac{2}{\sum_i \alpha_i^2}\right) = 1 - \frac{\alpha_i^2}{\sum_i \alpha_i^2}\cdot \tag{9}$$

Then define the statistic

$$\chi_{\text{pf}} = \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau})^2 = \sum_i r_i^2\cdot \tag{10}$$

The statistic χ_{pf} is similar to l^t and differs from l^t only in the use of $\hat{\tau}$ in place of τ. In addition, χ_{pf} is the sum of squares of estimated standardized residuals (r_i's) and can be

used to assess person fit for the LNMRT. If the LNMRT does not fit the response times of the examinee, some of the r_i's would be large in absolute value, and, as a consequence, χ_{pf} will be large. Appendix B includes R codes for computing χ_{pf} from a data set.

The Null Distribution of χ_{pf}

The statistic l^t can be expressed as

$$\begin{aligned} l^t &= \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau} - \hat{\tau} + \tau)^2 \\ &= \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau})^2 + (\hat{\tau} - \tau)^2 \sum_i \alpha_i^2 - 2(\hat{\tau} - \tau) \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau}) \qquad (11) \end{aligned}$$

The third term in the right-hand side of Equation 11 is

$$\begin{aligned} 2(\hat{\tau} - \tau) \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau}) &= 2(\hat{\tau} - \tau) \left[\sum_i \alpha_i^2(y_i - \beta_i) + \hat{\tau} \sum_i \alpha_i^2 \right] \\ &= 2(\hat{\tau} - \tau) \left[\sum_i \alpha_i^2(y_i - \beta_i) + \sum_i \alpha_i^2(\beta_i - y_i) \right], \qquad (12) \\ &= 0, \qquad (13) \end{aligned}$$

where the equality in Equation 12 holds because of Equation 2. Equations 11 and 13 imply that

$$\begin{aligned} l^t &= \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau})^2 + (\hat{\tau} - \tau)^2 \sum_i \alpha_i^2 \qquad (14) \\ &= \sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau})^2 + \frac{(\hat{\tau} - \tau)^2}{\text{Var}(\hat{\tau})}, \qquad (15) \end{aligned}$$

where the last equality holds because of Equation 3. Note that the first term in the right-hand side of Equation 15 is equal to χ_{pf}. Because $\hat{\tau}$ follows a normal distribution, the second term in the right-hand side of Equation 15 follows a χ^2 distribution with one degree of freedom (e.g., Rohatgi & Saleh, 2001, p. 324), that is,

$$\frac{(\hat{\tau} - \tau)^2}{\text{Var}(\hat{\tau})} \sim \chi_1^2. \qquad (16)$$

The covariance between $y_i - \beta_i + \hat{\tau}$ and $\hat{\tau} - \tau$ can be expressed as

$$\begin{aligned}
\mathrm{Cov}(y_i - \beta_i + \hat{\tau}, \hat{\tau} - \tau) &= \mathrm{Cov}(y_i - \beta_i + \hat{\tau}, \hat{\tau}) \\
&= \mathrm{Cov}(y_i, \hat{\tau}) + \mathrm{Var}(\hat{\tau}) \\
&= -\frac{1}{\sum_i \alpha_i^2} + \frac{1}{\sum_i \alpha_i^2} \qquad (17) \\
&= 0, \qquad (18)
\end{aligned}$$

where the equality in Equation 17 holds because of Equation 8. Because both $y_i - \beta_i + \hat{\tau}$ and $\hat{\tau} - \tau$ are linear functions of the y_i's and hence are normally distributed, Equation 18 implies that $(y_i - \beta_i + \hat{\tau})$ is independent of $(\hat{\tau} - \tau)$ because of Lemma 5.3.3 of Casella and Berger (2002, p. 220),[4] which implies that $\alpha_i^2(y_i - \beta_i + \hat{\tau})^2$ is independent of $\frac{(\hat{\tau}-\tau)^2}{\mathrm{Var}(\hat{\tau})}$. Therefore, the two terms in the right-hand side of Equation 15 are independent of each other. Then, Equations 5, 15, and 16 imply that l^t, which follows a χ_I^2 distribution, is a sum of two independent random variables $\sum_i \alpha_i^2(y_i - \beta_i + \hat{\tau})^2 = \chi_{\mathrm{pf}}$ and $\frac{(\hat{\tau}-\tau)^2}{\mathrm{Var}(\hat{\tau})}$, and the second of these variables follows a χ_1^2 distribution. Then, because the moment generating function (MGF) of a χ_I^2 random variable is given by (e.g., Rohatgi & Saleh, 2001, p. 215)

$$E(e^{u\chi_I^2}) = (1 - 2u)^{-I/2}, \qquad (19)$$

the MGF of l^t is given by

$$E(e^{ul^t}) = E\left(e^{u\left(\chi_{\mathrm{pf}} + \frac{(\hat{\tau}-\tau)^2}{\mathrm{Var}(\hat{\tau})}\right)}\right) = E\left(e^{u\chi_{\mathrm{pf}}}\right) E\left(e^{u\frac{(\hat{\tau}-\tau)^2}{\mathrm{Var}(\hat{\tau})}}\right), \qquad (20)$$

where the second equality holds because of the independence of the two terms in the right-hand side of Equation 15. Then, Equations 5, 16, 19, 20 imply that

$$(1 - 2u)^{-I/2} = E\left(e^{u\chi_{\mathrm{pf}}}\right)(1 - 2u)^{-1/2}, \qquad (21)$$

which implies that the MGF of χ_{pf} is given by

$$E\left(e^{u\chi_{\mathrm{pf}}}\right) = (1 - 2u)^{-(I-1)/2}, \qquad (22)$$

[4]The lemma states that two linear combinations of independent normal random variables are independent if their covariance is zero.

which is the MGF of a central χ^2 random variable with $(I-1)$ degree of freedom. By the uniqueness of MGFs (e.g., Rohatgi & Saleh, 2001, p. 88),

$$\chi_{\text{pf}} \sim \chi^2_{I-1}, \tag{23}$$

that is, χ_{pf} follows a central χ^2_{I-1} distribution under the null hypothesis of no misfit. Because no asymptotic/large-sample approximation was used in the derivation of the distribution of χ_{pf} under the null hypothesis, the number of items do not have to be large for the distributional result to hold.

If person misfit is present due to, for example, an examinee answering some items much quicker than expected and some other items much slower than expected, then χ_{pf} will be much larger than what is expected under a χ^2_{I-1} null distribution. If χ_{pf} is statistically significant for an examinee, an examination of the sr_i's of the examinee would reveal the items that were answered too soon or too late by the examinee.

Simulation

Two sets of simulations were performed. The first set of simulations, which involved analysis of data generated under no model misfit, was intended to compute the Type I error rates of l^t and χ_{pf}. The second set of simulations, which involved analysis of data generated under some misfit, was intended to compare the power of l^t and χ_{pf}. Test lengths of 20, 40, and 80 items were considered in each set of simulations. The number of examinees in a data set was found not to affect the Type I error rate or power of the person-fit statistics in preliminary simulations—so this number was fixed at 10,000.

Simulation of Data (with No Misfit) for Computing Type I Error Rates

Data were simulated under the LNMRT given by Equation 1. The true values of α_i's and β_i's were simulated from a $\mathcal{N}(1.87, 0.15^2)$ and a $\mathcal{N}(4, 0.45^2)$ distribution, respectively. The true values of τ_j's were simulated from a $\mathcal{N}(0, 0.3^2)$ distribution. These generating distributions were selected to make the summary of the simulated data look like the

summary of the real data described in van der Linden (2006).[5] For each test length, a total of 1,000 data sets were used to compute the Type I error rates of the person-fit statistics. A new set of true values of the parameters was used to simulate each of these 1,000 data sets. For each simulated data set, the item parameter estimates (posterior means) were computed using the R package LNIRT (Fox et al., 2017)[6] and then these estimates were used to compute the person-fit statistics. The computation of l^t for each simulated data set was based on a sample of size 5000 drawn from the posterior distribution of τ.[7] For a test length, the Type I error rate of a person-fit statistic was computed as the proportion of examinees that had a statistically significant value of that statistic for that test length.

Simulation of Data (with Some Misfit) for Computing Power

In this set of simulations, the generated data sets included 90% examinees whose response times followed the LNMRT given by Equation 1—the response times of these examinees were simulated in a manner similar to that for data simulated under no model misfit—and 10% aberrant or misfitting examinees whose response times did not follow the LNMRT. Two types of person misfit were considered, one due to item preknowledge (so that some items were assumed to be compromised) and another due to random responding. The former type of misfit was considered in van der Linden and Guo (2008) and the latter by Marianti et al. (2014).

To create person misfit due to item preknowledge, it was assumed that the response-times of the aberrant examinees followed the LNMRT given by Equation 1 for the non-compromised items, but were equal to 5, 10, 15, or 20 seconds for the compromised

[5]For example, with these generating distributions, the mean response times of the items were roughly between 25 and 150 seconds and the mean response times of the persons were between 20 and 170 seconds; these values roughly match the corresponding quantities in Figures 1 and 2 of van der Linden (2006).

[6]The option "WL=1" was used to ensure that the model provided in Equation 1 is fitted by the package.

[7]Equations 14 and 15 of van der Linden (2006) provide the mean and variance of the posterior distribution of τ that is a normal distribution. To draw a sample from this posterior distribution, the standard procedure of drawing normal random variables was used.

items. The percent of compromised items on the test was assumed to be 5, 10, 20, or 30. The set of compromised items was assumed to be the same for all examinees in a data set. To create person misfit due to random responding, it was assumed that the response-times of the aberrant examinees followed the LNMRT given by Equation 1 for most items; however, the response-times of the aberrant examinees to 10%, 30%, or 50% items were simulated from a lognormal distribution whose mean is given by Equation 1 but whose standard deviation (SD) is two, three, or four times that given by Equation 1.

For each simulation condition represented by a test length, a percent of aberrant examinees, and a specific type of aberrance,[8] the following steps were iterated 1,000 times:

1. Simulate a data set with 90% non-aberrant examinees and 10% aberrant examinees
2. Estimate the item parameters of the data set
3. Compute the MLEs of the person speed parameters for the data set using the item-parameters estimated in the previous step.[9] Draw a sample of size 5,000 from the posterior distribution of τ of each examinee using the item-parameters estimated in the previous step
4. Compute the two person-fit statistics for all examinees in the data set using the item and person parameter estimates and the posterior samples obtained in the previous two steps

The power of each person-fit statistic for each simulation condition was computed as the proportion of aberrant examinees that had a statistically significant value of the statistic under that simulation condition.

[8]For example, for item preknowledge, a specific type of aberrance refers to a specific percentage of items that were compromised and a specific time that the aberrant examinees took to answer the compromised items.

[9]Another set of calculations were done in which item parameters were not estimated and fixed at their true values—these calculations produced results (not shown here and can be obtained from the authors upon request) that are similar to those reported in this paper.

Results for Data Simulated Under No Misfit and the Type I Error Rates

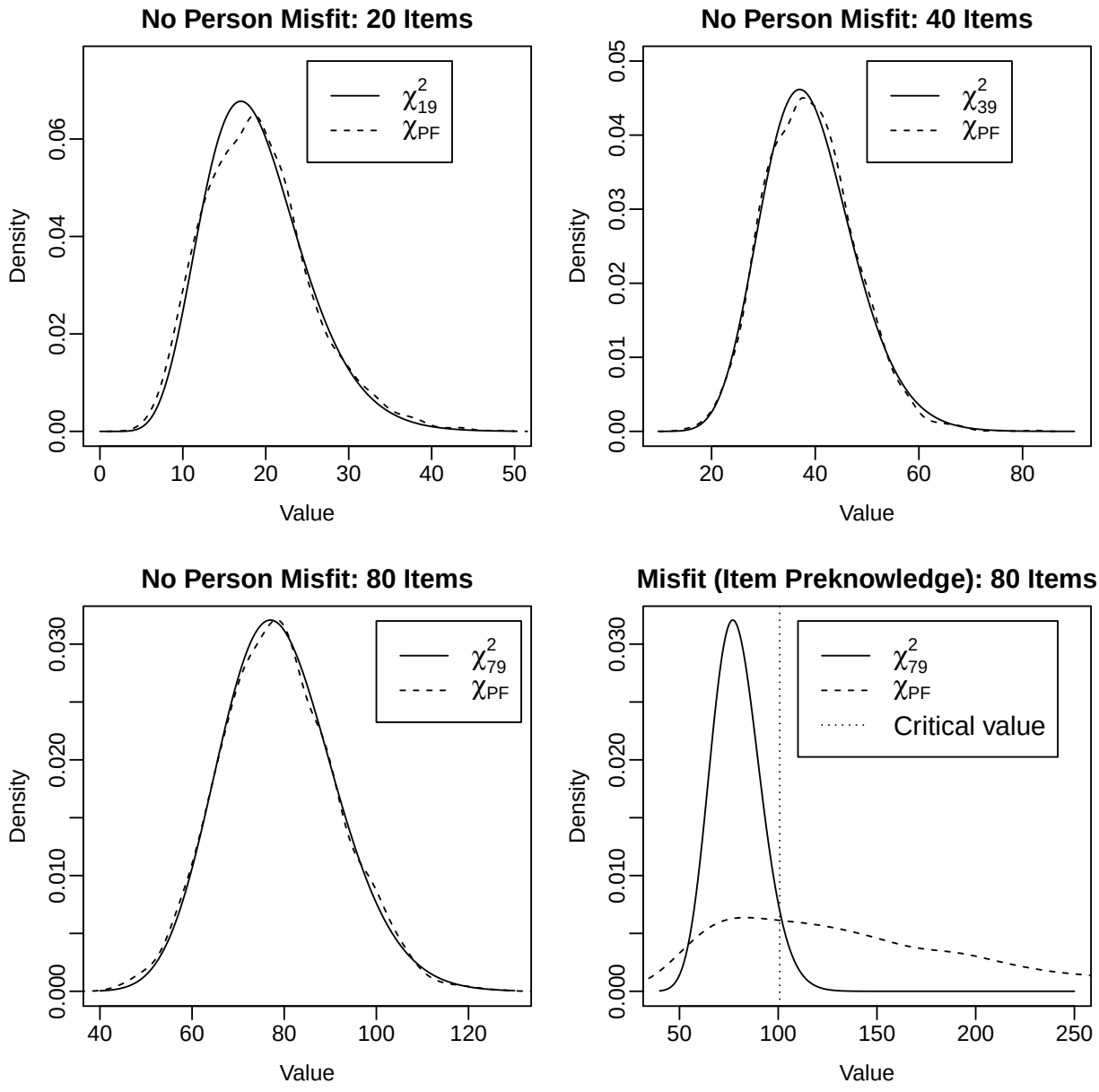

Figure 1: The distribution of χ_{pf} under no misfit (the first three panels) and under item preknowledge (bottom right panel).

The top two panels and the bottom left panel of Figure 1 show the kernel-density

Table 1: Type I Error Rates

Statistic	20 Items	40 Items	80 Items
l^t	0.036	0.040	0.043
χ_{pf}	0.050	0.050	0.050

estimates[10] of the distributions of the values of χ_{pf} for random subsets of 2,000 simulated non-aberrant examinees for 20 items, 40 items, and 80 items, respectively. The theorized null distributions of χ_{pf} for the three test lengths (χ^2_{19}, χ^2_{49}, and χ^2_{79}, respectively) are also shown for convenience. The ranges of the horizontal and vertical axes are different in the four panels in the figure. The distribution of the values of the χ_{pf} statistic is very close, especially at the right tail, to the corresponding theorized null distribution for each test length in the three panels. There is some difference between the values of the χ_{pf} and the theorized null distributions, especially near the peak of the distributions, but the magnitude of the difference decreases as test length increases. Thus, the null distribution shown in Equation 23 seems to hold for data simulated under no misfit.

Table 1 shows the Type I error rates at 5% level (rounded to three decimal places) of the two person-fit statistics. The table shows that l^t is slightly conservative, that is, its Type I error rate is slightly smaller than the nominal level of 0.05, although the extent of conservativeness decreases as test length increases; the Type I error rate of χ_{pf} is very close to the nominal level. The Type I error rates at 1% level (not shown here and can be obtained from the authors upon request) led to similar conclusions.

Results for Data Simulated Under Some Misfit and Power

The bottom right panel of Figure 1 shows the kernel-density estimates of the distribution of the values of χ_{pf} for a random subset of 2,000 simulated aberrant examinees who had item preknowledge on 32 items and answered them in 15 seconds on 80-item tests. The corresponding theorized null distribution (χ^2_{79}) of χ_{pf} and the critical value at 5% level (or, the 95th percentile of the χ^2_{79} distribution) are also shown for convenience. The

[10]The figure was created using the function "density" in the R software (R Core Team, 2017).

figure shows that the values of χ_{pf} under person misfit (due to item preknowledge) are much larger on average compared to what is expected under no misfit.[11]

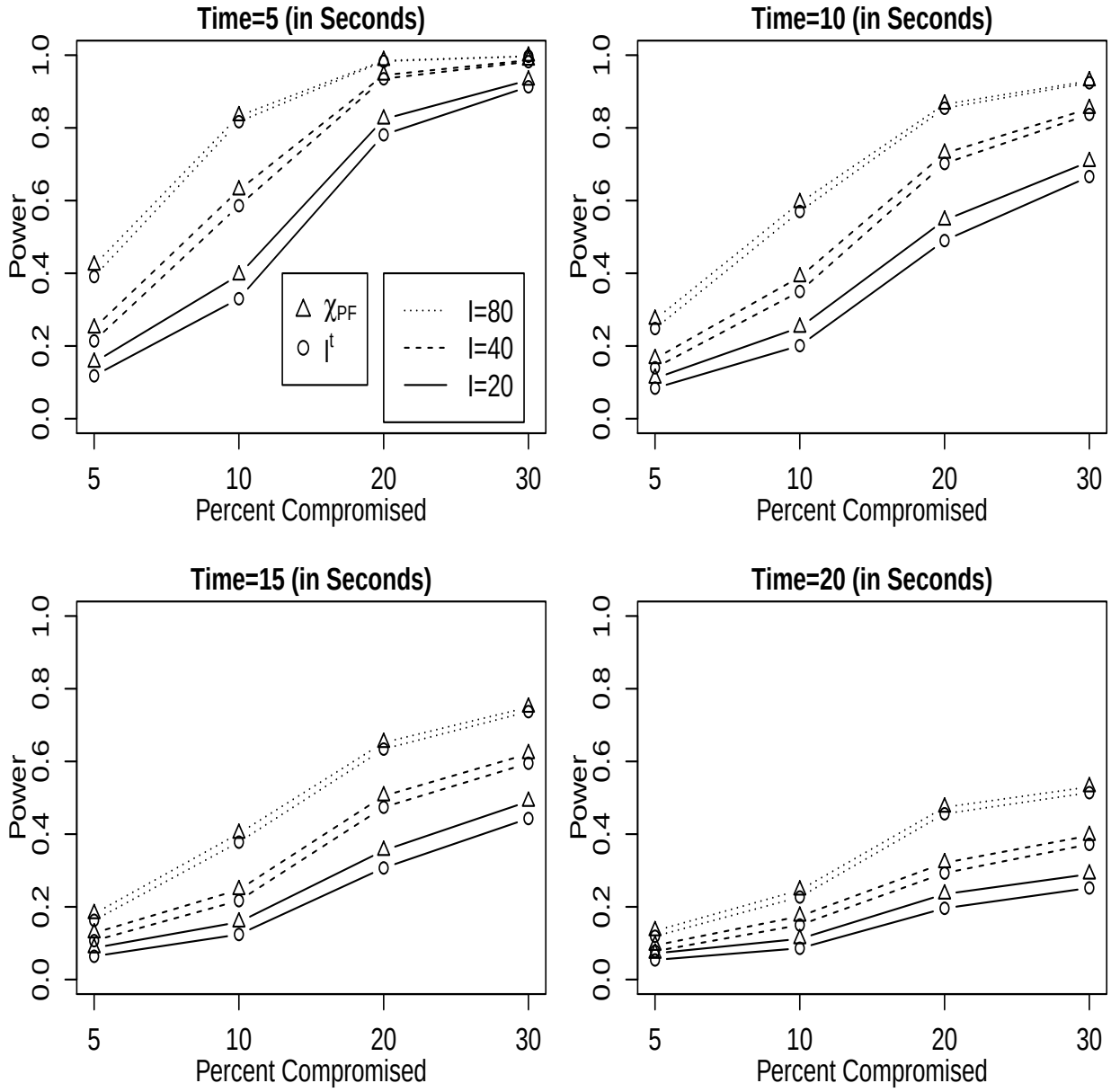

Figure 2: Power at 5% Level of l^t and χ_{pf} to Detect Person Misfit Due to Item Preknowledge.

Figure 2 shows the average power to detect person misfit due to item preknowledge for different values of test length, percent of items compromised, and time-taken-to-answer-the-

[11]Roughly 60% of the values of χ_{pf} are larger than the critical value (that is, the power is about 0.60).

compromised-items of l^t and χ_{pf}. The values of power at 5% level are reported in Figure 2. The four panels of the figure show the average power of the statistics when the time taken to answer the compromised items was 5, 10, 15, and 20 seconds, respectively; the time is shown in the title of each panel. The percent of the items that are compromised is shown along the X-axis and the power is shown along the Y-axis. The two solid lines show the average power for test length (I) of 20, the two dashed lines are for test length of 40, and the two dotted lines are for test length of 80. The power for χ_{pf} and l^t are shown using hollow triangles and hollow circles, respectively, joined by a line.

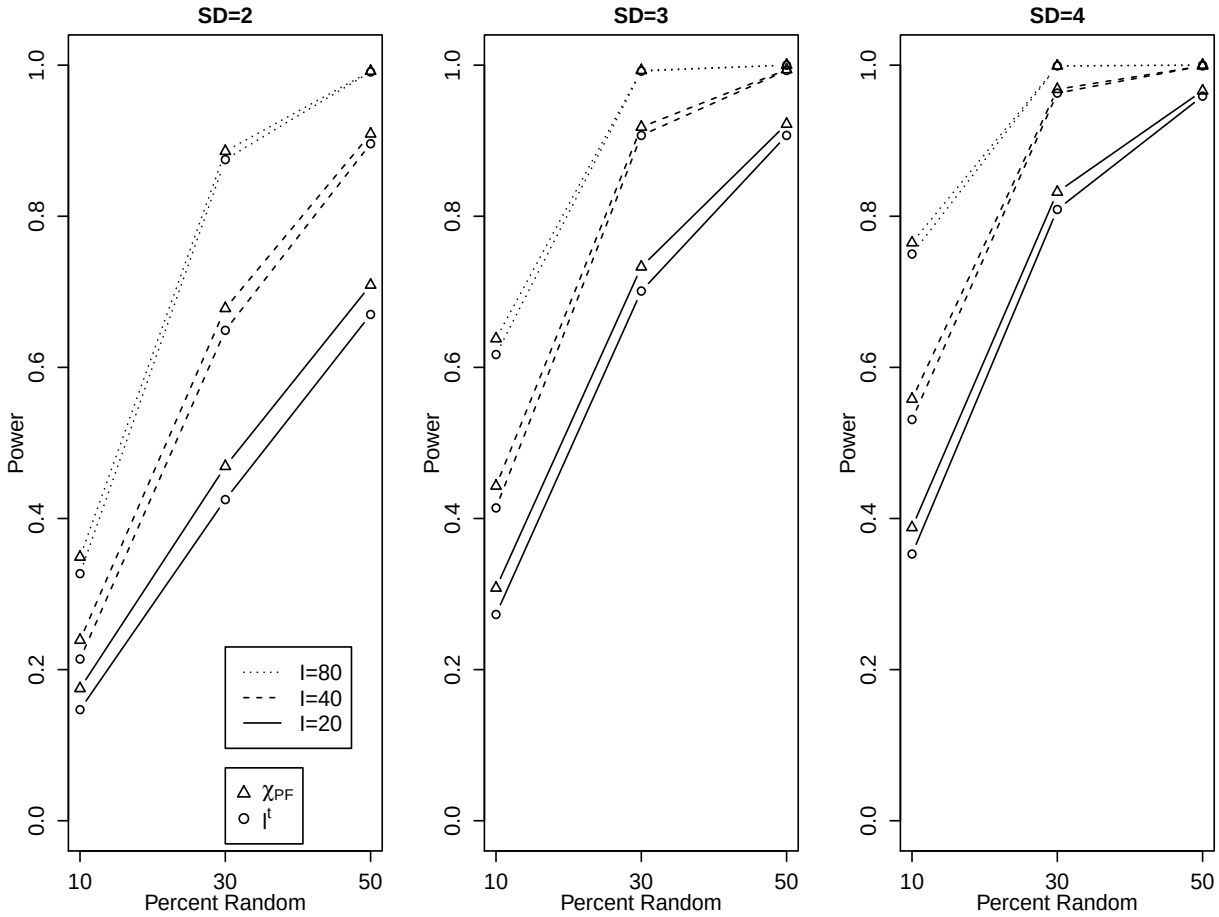

Figure 3: Power at 5% Level of l^t and χ_{pf} to Detect Person Misfit Due to Random Responding.

Figure 3 shows the average power at 5% level to detect person misfit due to random

responding for different values of test length, percent of items on which aberrant examinees responded randomly, and the SD of the generating distribution (as a multiple of the SD of the distribution shown in Equation 1) for random responding of the two person-fit statistics. The three panels of the figure show the average power of the statistics when the SD of the generating distribution for random responding was respectively 2, 3, and 4 times the SD of the distribution shown in Equation 1. The percent of items on which aberrant examinees responded randomly is shown along the X-axis and the power is shown along the Y-axis.

Figures 2 and 3 shows that:

- Power increases with test length, which is a favorable result for the person-fit statistics.
- Power becomes larger as the extremeness of the response times of the aberrant examinees increases. For example, in Figure 2, the power in the top left panel (for 5 seconds) is larger compared to the other panels and, in Figure 3, the power in the rightmost panel (for SD=4) is larger compared to the other panels.
- Power increases as the percent of items with aberrant responding increases (this is reflected by an increase in power of each statistic from the left to the right in each panel).
- The statistic χ_{pf} is always slightly more powerful compared to l^t. The difference is noticeable under several conditions, for example, for 20-item tests in Figure 2 where the difference is as large as 0.05 in a few cases.[12]

Two figures showing the power of the statistics at 1% level of significance are shown in Appendix A—those two plots appear similar to Figures 2 and 3 except that the power for 20 items is considerably smaller at 1% level compared to 5% level.

Real Data Example

Let us consider a real data set that consists of the responses and response times of more than 18,000 test takers on a computerized test for English proficiency. The test includes

[12]This result is in agreement with the conservative Type I error rate of l^t in Table 1.

34 operational items that are all multiple-choice—data on only these items are analyzed here. The number of non-operational items administered to each examinee varies and the time allocated to each examinee varies according to the number of non-operational items. The mean response times on the operational items were between 21 and 52 seconds and the mean response times of the persons on the operational items were between 9 and 53 seconds.

The posterior means of the item parameters for the LNMRT were computed for the data set using the R package LNIRT (Fox et al., 2017). The standardized residuals of van der Linden and Guo (2008) were statistically significant for 4.65% examinee-item combinations at 5% level, which indicates that the LNMRT is a reasonable model for these data. The posterior means of the item parameters were used as the estimated item parameters in the computation of χ_{pf} and l^t. At 5% level of significance, person misfit was found for 10.5% examinees using l^t and for 11.6% examinees using χ_{pf}—a larger percentage for χ_{pf} is in agreement with results in the simulation study. The set of 11.6% examinees for which misfit was found using χ_{pf} included all but two of the examinees for whom misfit was found using l^t as well as 1.1% for whom misfit was not found using l^t. At 1% level of significance, person misfit was found for 5.4% examinees using l^t and for 6.2% examinees using χ_{pf}.

The two panels of Figure 4, like Figures 3-5 of van der Linden and Guo (2008), show the standardized residuals sr_i's for all the items for two examinees for whom misfit was found using both χ_{pf} and l^t. Note that the sr_i's are roughly distributed as standard normal variables if the LNMRT fits the data (van der Linden & Guo, 2008). In each panel, the item number is shown along the horizontal axis and the sr_i's are shown using a bar along the vertical axis. Positive sr_i's are indicated by a bar above the 0-line and negative sr_i's are indicated by a bar below the 0-line. A bar is hollow if the corresponding answer is incorrect and filled in black if the corresponding answer is correct. Horizontal (dotted) lines are shown at 2 and -2 for convenience of evaluating whether a specific sr_i is statistically significant. For the examinee represented in the top panel, the response time was 110 seconds for Item 22; consequently, sr_i was large for the item; the examinee took longer than

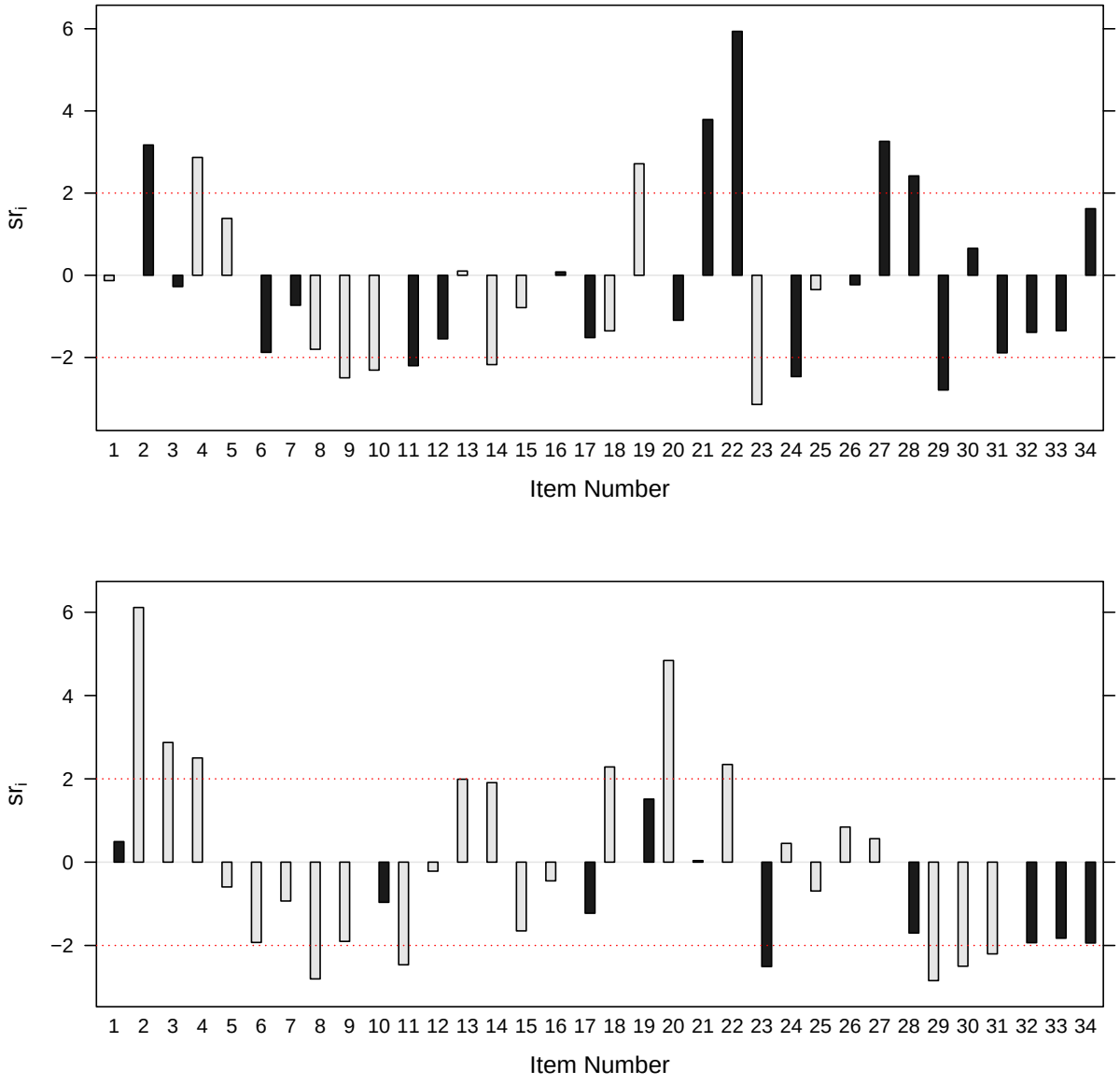

Figure 4: Plots of the sr_i's for all the items for two examinees

expected on a few other items such as Items 2 and 21. For the examinee represented in the bottom panel, the response times were much longer than expected for Items 2-4 (especially, the examinee took about four minutes to answer Item 2) and shorted than expected on the last seven items.

Conclusions and Recommendations

This paper is the first to perform frequentist person-fit analysis based on response times. Specifically, this paper focuses on the log-normal response-time model (van der Linden, 2006) and suggests a new person-fit statistic for the model. Thus, this paper promises to be a significant step in completing a frequentist toolkit for assessing fit of response-time models that included statistics for assessing other types of fit (e.g., Glas & van der Linden, 2010; Ranger & Ortner, 2012; van der Linden & Glas, 2010), but did not include any person-fit statistic until now. The null distribution of the suggested statistic is proved to be a χ^2 distribution. A simulation study demonstrates that the Type I error rate of the new statistic is very close to the nominal level and the power of the statistic is slightly larger than that of an existing person-fit statistic (l^t suggested by Marianti et al., 2014). A real data application of the new statistic is also included. Computer code for computing the statistic is provided. The new statistic can be interpreted as the sum of squares of estimated standardized residuals and can be calculated very easily, as is clear from the computer code—so the statistic promises to be useful to those interested in response-time models.

One important question is "How should the new person-fit statistic be used in practice?" In the extensive literature on person-fit statistics based on item scores, experts mostly recommend the use of person-fit statistics when an investigator wants to test against an unspecified general alternative (e.g., Glas & Dagohoy, 2007; Sinharay, 2016)—the same recommendation applies here. If the anticipated model violation is more specific (such as speededness), other statistics may be more appropriate. Researchers such as van der Linden and Guo (2008) suggested employing person-fit statistics based on response times to detect aberrant examinee behavior as a part of quality control and the χ_{pf} statistic can be used in the same way. However, van der Linden and Guo (2008) warned practitioners against the mechanical use of person-fit statistics based on response times in high-stakes contexts such as detection of cheating because aberrant response-time patterns may arise due to bad time management. A prudent strategy in high-stakes contexts would involve the use of χ_{pf} and/or other statistics as secondary evidence, as recommended by researchers such as

Holland (1996) and Hanson, Harris, and Brennan (1987). For example, the bottom panel of Figure 4 does not provide convincing evidence on its own that the corresponding examinee cheated; however, if a proctor report shows that the examinee may have copied the answers to the last few items from another examinee, then the panel provides much more convincing evidence of cheating.

The choice of the significance level in applications of person-fit statistics is a crucial issue. Person-fit statistics using response time have been recommended for detection of test fraud (e.g. van der Linden & Guo, 2008) and Wollack, Cohen, and Eckerly (2015) commented that methods for detection of test fraud are typically applied with conservative levels. However, an encouraging aspect of the the new person-fit statistic is that the statistic appears to have satisfactory power in several cases (see Figures A1 and A2) even for the conservative significance level of 1%.

In the context of person-fit analysis using item scores, researchers such as Meijer and Tendeiro (2012) and Sinharay (2017) emphasized the importance of first assessing the overall fit of the IRT model before performing any person-fit analysis. A similar strategy should be used for person-fit analysis using response times. In the context of this paper, a similar strategy consists of assessing the overall fit of the LNMRT to the data set before applying the suggested χ_{pf} statistic. If the LNMRT does not fit the data overall (due to, for example, a violation of the local independence assumption), then the null distribution of χ_{pf} may not be χ^2 and the use of the statistic could lead to erroneous conclusions. Glas and van der Linden (2010) suggested several Lagrangian Multiplier statistics for assessing the fit of the LNMRT in general and Schnipke and Scrams (1999) used graphical plots and the root mean squared error between the observed and predicted cumulative distribution function to assess the fit of the LNMRT in general;[13] one or more of these methods may be used to ensure the good fit of the LNMRT before computing χ_{pf} for the data. A residual-based model-fit analysis method outlined in van der Linden and Guo (2008) revealed an adequate fit of the LNMRT to the data from the English proficiency test that was used earlier in this

[13]Schnipke and Scrams (1999) found the LNMRT to be the best-fitting model among four models for data from the Law School Admission Test.

paper—so the application of the person-fit analysis to that example seems reasonable.

This paper has several limitations and, consequently, leaves room for plenty of future research. First, the suggested statistic should be computed for more simulated and real data sets. Second, only the LNMRT was considered in this paper—extension of the suggested statistic to other types of response-time models would be a potential area of future research. It is anticipated that for other response-time models, the suggested person-fit statistics would have a standard-normal distribution under the null hypothesis only for long tests. Third, research on finding an appropriate effect size for the suggested person-fit statistic would be useful. Fourth, it is possible to examine the consequences of misfit of the LNMRT on the properties of the new person-fit statistic in future research. Finally, item parameters were assumed known in the person-fit analysis in this paper; the effect of this assumption on the properties of the person-fit statistics and approaches for accounting for the uncertainty of the item-parameters in the distribution of the new person-fit statistic (possibly using a Bayesian approach) may be studied in future research.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis, 2nd edition.* New York, NY: Wiley.

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13–38.

Boughton, K., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 177–190). Washington, DC: Routledge.

Casella, G., & Berger, R. L. (2002). *Statistical inference (2nd edition).* Pacific Grove, CA: Duxbury.

Fox, J.-P., Entink, R. K., & Klotzke, K. (2017). *LNIRT: Lognormal response time item response theory models.* (R package version 0.2.0)

Fox, J.-P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, *54*, 243–262.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York, NY: Chapman and Hall.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*, 159–180.

Glas, C. A. W., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology*, *63*, 603–626.

Hanson, B. A., Harris, D. J., & Brennan, R. L. (1987). *A comparison of several statistical methods for examining allegations of copying (ACT research report series no. 87-15).* Iowa City, IA: American College Testing.

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (ETS Research Report No. RR-94-4). Princeton, NJ: ETS.

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48.

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, *62*, 621–640.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*, 359–379.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*, 426–451.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, *58*, 445–469.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Meijer, R. R., & Tendeiro, J. N. (2012). The use of the l_z and l_z^* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics*, *37*, 758–766.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna,

Austria.

Ranger, J., & Ortner, T. (2012). The case of dependency of responses and response times: A modeling approach based on standard latent trait models. *Psychological Test and Assessment Modeling*, *54*, 128–148.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rohatgi, V. K., & Saleh, A. K. M. E. (2001). *An introduction to probability and statistics.* New York, NY: Wiley.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*, 213–232.

Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks* (LSAC-R-97-09). Newtown, PA: Law School Admission Council.

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ, US: Lawrence Erlbaum Associates.

Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, *81*, 992–1013.

Sinharay, S. (2017). How to compare parametric and nonparametric person-fit statistics with real data. *Journal of Educational Measurement*, *54*, 420–439.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person

parameter. *Psychometrika*, *66*, 331–342.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 179–203). New York, NY: Academic Press.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272.

van der Linden, W. J. (2016). Lognormal response-time model. In W. van der Linden (Ed.), *Handbook of item response theory, Volume 1. Models.* Boca Raton, FL: Chapman and Hall/CRC.

van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, *75*, 120–139.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339–356.

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy

and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*, 317–345.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339.

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. *Educational and Psychological Measurement*, *75*, 931–953.

Appendix A: Power for 1% Level of Significance

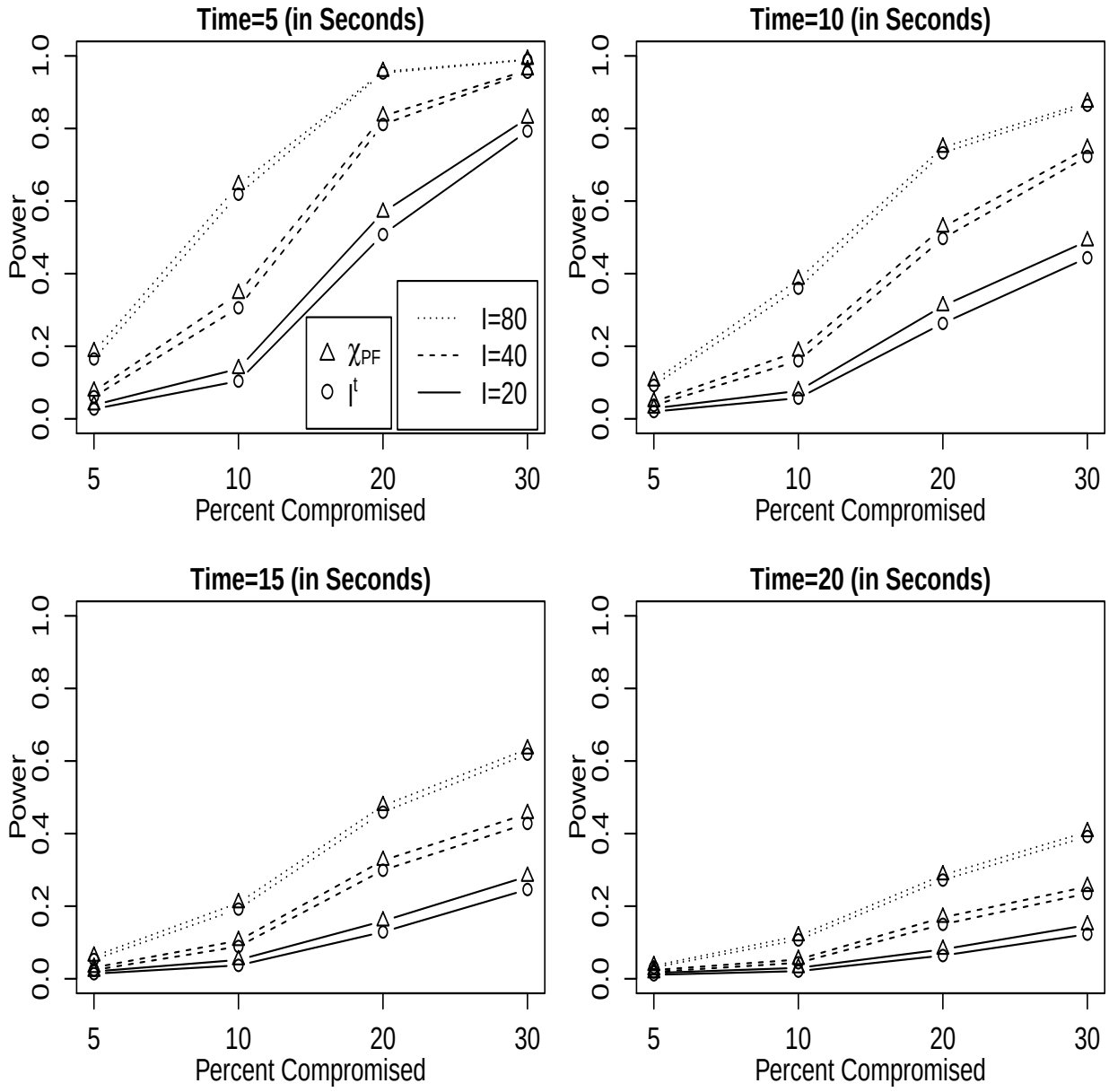

Figure A1: Power at 1% Level of l^t and χ_{pf} to Detect Person Misfit Due to Item Preknowledge.

Figure A1 shows the average power at 1% level to detect person misfit due to item preknowledge. Figure A2 shows the average power at 1% level to detect person misfit due

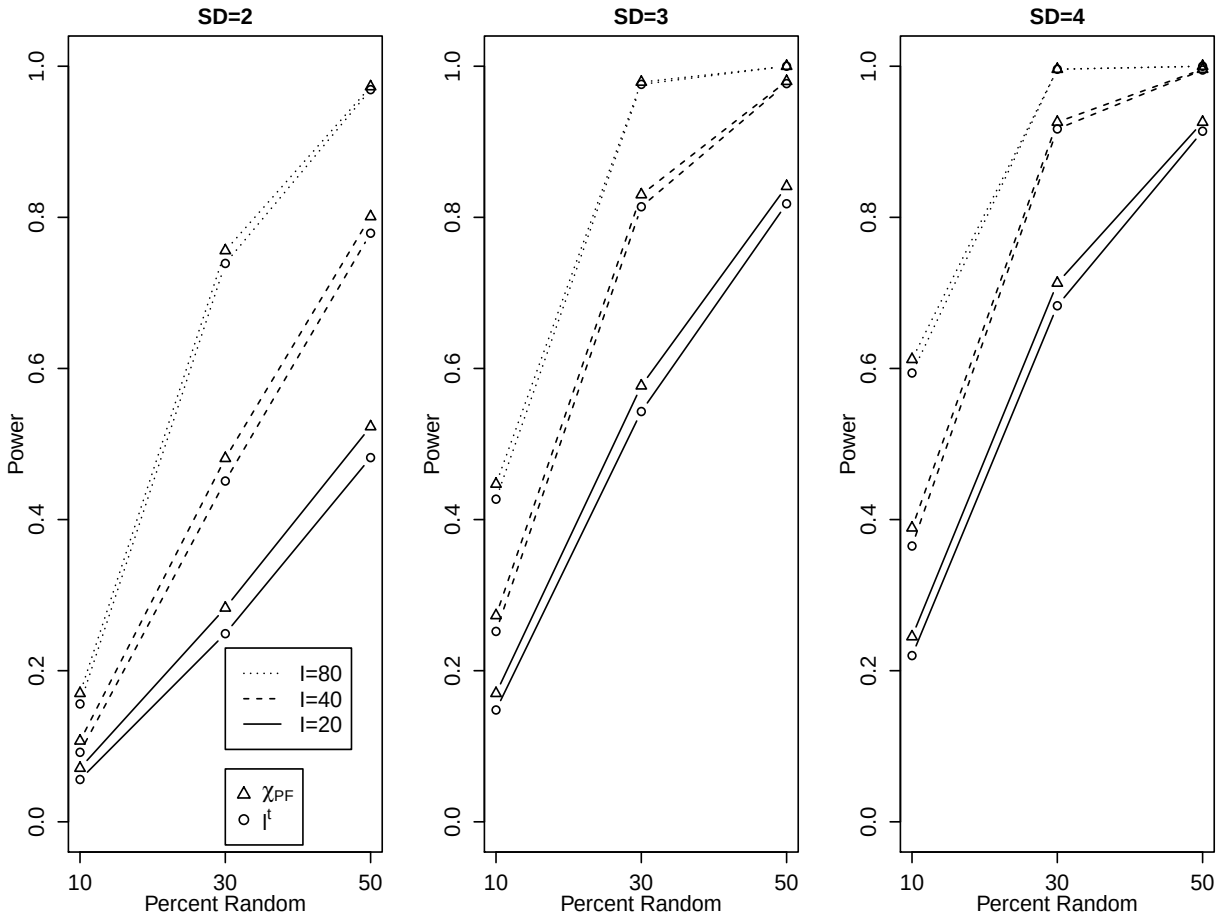

Figure A2: Power at 1% Level of l^t and χ_{pf} to Detect Person Misfit Due to Random Responding.

to random responding. The values of power at 1% level are smaller than those at 5% level, but the patterns including the relative performace of l^t and χ_{pf} at 1% level (represented in Figures A1 and A2) are very similar to those at 5% level (represented in Figures 2 and 3).

Appendix B: R Subroutine to Compute the New Person-fit Statistic

```
# R Subroutine to Compute the New Person-fit Statistic given the Item Parameters
# (alpha and beta) and the Data Set 'ltimes' that Includes the Log-response Times
ChiPF=function(alpha,beta,ltimes)
{n=nrow(ltimes)#n is the number of examinees in the data file
 tauhat=rep(sum(alpha*alpha*beta)/sum(alpha*alpha),n)-ltimes%*%(alpha*alpha)
        /(sum(alpha*alpha))#tauhat's are the estimated examinee parameters
 v=ltimes-rep(1,n)%*%t(beta)+tauhat%*%t(rep(1,length(beta)))
 chisq=(v*v)%*%(alpha*alpha)
 return(chisq)# 'chisq' contains the values of the new statistic for all examinees
}
```